In addition to scatter plots, we can create several other types of plots in R. Two examples are histograms and box plots.

Let's first create a histogram of CellularSubscribers.

To do this, we'll use the hist function.

So in your R console type hist, and then in parentheses WHO$CellularSubscribers.

Close the parentheses and hit Enter.

If you go over to your plotting window you can see that the values of CellularSubscribers are shown on the x-axis and the frequency of these values is shown on the y-axis.

A histogram is useful for understanding the distribution of a variable.

Here we can see that the most frequent value of CellularSubscribers is around 100.

We can also easily create a box plot in R.

We'll make a box plot of LifeExpectancy sorted by Region.

So back in your R console type boxplot, and then in parentheses, WHO$LifeExpectancy and then a tilde symbol followed by WHO$Region.

Close the parentheses and hit Enter.

Then go over to your plotting window.

You may need to stretch it out a little bit so that you can see all of the labels on the x-axis.

A box plot is useful for understanding the statistical range of a variable.

This box plot shows how life expectancy in countries varies according to the region the country is in.

The box for each region shows the range between the first and third quartiles with the middle line marking the median value.

The dashed lines at the top and bottom of the box, often called whiskers, show the range from the minimum to maximum values, excluding any outliers, which are plotted as circles.

Outliers are defined by first computing the difference between the first and third quartiles, or the height of the box.

This number is called the inter-quartile range.

Any point that is greater than the third quartile plus the inter-quartile range, or any point that is less than the first quartile minus the inter-quartile range is considered an outlier.

This box plot shows us that Europe has the highest median life expectancy, the Americas has the smallest inter-quartile range, and the eastern Mediterranean region has the highest overall range of life expectancy values.

If you want to give nice labels to any of your plots, you can easily do so by adding a few arguments.

Go back to your R console, scroll up and then inside the parentheses type a comma and then xlab equals and then empty quotes-- we're not going to label the x-axis here because the regions are already nicely labeled-- and then a comma, and then ylab equals "Life Expectancy".

Close the quotes, and then a comma, and then main = "Life Expectancy of Countries by Region".

Close the quotes and hit Enter.

If you go back and look at your box plot again you should now see that there's a nice y-axis label and an overall title to the plot.

Lastly, let's take a look at some summary tables.

So go back to your R console and we'll start by making a table of the Region variable.

So we'll type table and then in parentheses WHO$Region.

Close the parentheses and hit Enter.

This is similar to what we saw in the summary output and counts the number of observations in each category of Region.

Tables work well for variables with only a few possible values, and we'll see more of this in recitation.

You can see some nice information about numerical variables by using the tapply function.

Let's start by looking at an example.

So type tapply, and then in parentheses WHO$Over60 comma, and then WHO$Region comma, and then mean.

Close the parentheses and hit Enter.

This splits the observations by Region and then computes the mean of the variable Over60.

So tapply splits the data by the second argument you give, and then applies the third argument function to the variable given as the first argument.

This result tells us that the average percentage of the population over 60 in African countries is about 5%, while the average percentage of the population over 60 in European countries is about 20%.

Let's look at another example.

This time in the tapply function, we'll give as the first argument WHO$LiteracyRate then as the second argument we'll give WHO$Region again.

And as our third argument we'll give min.

Close the parentheses and hit Enter.

Here we see something a little strange.

We have the value NA for all of the regions.

This is because we have some missing values in our data for literacy rate.

A common thing to do is to just remove the missing values when doing the computation.

We need to pass one additional argument, so hit the up arrow, and then inside the parentheses add a comma and then na.rm = TRUE and hit Enter.

This removes all of the countries that are missing a value for LiteracyRate before doing the computation.

This time we see numerical values, as we expect.

So we've split the data by Region again and computed the minimum value of LiteracyRate for all countries with a value in the LiteracyRate variable.

By using some basic functions in R, plots, and summary tables we were able to get a better understanding of our data.

You'll see more of this in the recitation and homework assignment.